Report on Bachelor Thesis

Institute of Economic Studies, Faculty of Social Sciences, Charles University in Prague

Student:	**Tomáš Turlík**
Advisor:	**doc. PhDr. Ladislav Krištoufek Ph.D.**
Title of the thesis:	**Neural networks and tree-based credit scoring models**

OVERALL ASSESSMENT:

The thesis submitted by Tomáš Turlík assesses the performance of machine learning techniques on loan default prediction and aims to answer the question whether these methods can be used instead or alongside logistic regressions. After a brief but sufficient introduction into credit scoring models, Mr. Turlík presents a well-developed review of machine learning methods which he then uses to challenge the commonly used approaches. The literature review thus provides a solid basis for the models used. For the analysis Mr. Turlík used the German Credit dataset and data provided by the Lending Club, a US lending platform. The datasets contrast in size since the former contains 1000 of observations and the latter almost 450 000 observations.

In the empirical part the study shows that on the selected datasets neither of reviewed models showed any greater significant superior predictive power. The models are clearly described and Mr. Turlík demonstrates his grasp of the methods used. Despite the different datasets, the models performed similarly in both cases. The thesis thus provides evidence that machine learning techniques achieve comparable predictive power and I therefore recommend the thesis for defence and suggest the grade A.

Suggested questions for defence:

In the Conclusion (p. 35), you mention that "*Given the time requirements for training and tuning the networks, there could plausibly exist cases when logistic regression is preferred to the neural networks or stacked ensembles even if it is worse"*. You also mention that the best model showed a "*relatively minor improvement*" (in case of the German model, p. 33) in the ROC AUC. These statements seem to prioritise the usage of the logistic regression. Could you please elaborate further the conclusion? In which cases would you prefer the logistic regression?

The literature review mentions Support Vector Machine (SVM) as a commonly used technique. Why this technique was not considered despite its widespread usage?

The data used for the analysis likely suffer from selection bias. The Conclusion (p. 36) mentions: ''*Underlying economic factors together with the selection bias problem, should also be incorporated into the scoring process.*'' How should the selection bias issue be addressed?

With regards to the limited size of the German dataset (which has only 1000 observation) isn't even 40 neurons likely to cause overfitting issues? Shouldn't the dataset be much larger to be analyzed by a neural network?

Report on Bachelor Thesis

Institute of Economic Studies, Faculty of Social Sciences, Charles University in Prague

Student:	**Tomáš Turlík**
Advisor:	**doc. PhDr. Ladislav Krištoufek Ph.D.**
Title of the thesis:	**Neural networks and tree-based credit scoring models**

***SUMMARY OF POINTS AWARDED**:*

CATEGORY	**POINTS**
Contribution (max. 30 points)	30
Methods (max. 30 points)	27
Literature (max. 20 points)	18
Manuscript Form (max. 20 points)	16
TOTAL POINTS *(max. 100 points)*	**91**
GRADE (A – B – C – D – E – F)	**A**

NAME OF THE REFEREE: Nicolas Fanta

DATE OF EVALUATION: 28th of August 2018

Referee Signature